A load-balanced optical packet switch architecture with an *O(1)* scheduling complexity

A. Goulet [(1)], A. Cassinelli [(1)], M. Naruse [(2)], F. Kubota [(2)], M. Ishikawa [(1)]
(1): University of Tokyo, Dept. Information Physics and Computing, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan,
Tel. +81-3-5841-6937 ; e-mail: {Alain_Goulet, Alvaro_Cassinelli, Masatoshi_Ishikawa}@ipc.i.u-tokyo.ac.jp
(2): National Institute of Information and Communications Technology, 4-2-1 Nukui-kita, Koganei, Tokyo 184-8795, Japan,
Tel. +81-42-327-6209 ; e-mail : {naruse, kubota}@nict.go.jp

Abstract: *A photonic implementation of the load-balanced switch architecture is presented. Because connection patterns are known and periodic, packet transmission can be pre-determined, making optical buffering practical.*

I - Introduction

Packet switching in core and edge routers currently relies on electronic switching fabrics, which necessitates expensive optical-electrical-optical conversions. In addition, single-stage switch architectures with centralized scheduling are becoming increasingly impractical as the number of channels and the line rate become higher [1]. These reasons amply justify research on all-optical packet switches.

A major drawback of photonics, however, is the lack of optical random access memories. Only optical buffers (OBs) using fiber delay lines (FDLs) are feasible at present [2]. The load-balanced switch architecture presented here uses regular connection patterns, allowing calculation of packet delays. Therefore, it makes efficient use of OBs. Moreover, no centralized scheduler is needed.

In the following, it is assumed that time is slotted and synchronized, at most one packet arrives per time slot, and the number of inputs/outputs is *N*.

II - Review of the load-balanced switch architecture

The load-balanced switch architecture proposed by C.S. Chang et al. [3] consists of two switch stages and one buffer stage (see Fig. 1). The first switch performs load-balancing; it makes bursty input traffic uniformly distributed. A virtual output queue (VOQ) composed of *N* separate FIFO queues stores the arriving packets at each port of the buffer stage: a packet with destination *j* is placed in queue *j*. The second switch performs switching. Because traffic is uniformly distributed, a deterministic TDM-like schedule that serves an input to an output $1/N^{th}$ of the time gives a 100% throughput if the traffic is weakly mixed [3].

This architecture provides additional advantages: it makes use of buffers efficiently, gives low average delay under heavy load and bursty traffic, and is scalable [3]. However, it does not preserve packet order.

III - Load-balanced optical packet switch (LB-OPS)

1- Load-balancing and TDM switches.

The load-balancing and TDM switches are in fact alike. They can be emulated by a switch that achieves full access (i.e., any input can be connected to any output) through a periodic sequence of *N* permutations.

A simple architecture to do so is the stage-controlled banyan network (SC-BN). It is a $\log_2 N$ multistage interconnection

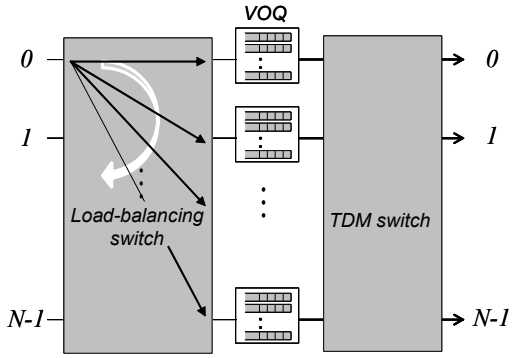

Fig. 1. Load-balanced switch architecture.

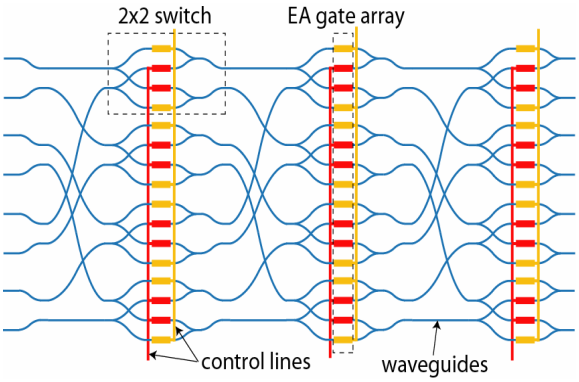

Fig. 2. 8×8 SC-BN with EA gates.

network formed from 2×2 switches. All the *N/2* switches within a stage are set either in the bar state or cross state; hence only $\log_2 N$ control signals are necessary to operate the switch. It has been proven that the set of *N* permutations obtained from an SC-BN provides full access [4].

Owing to their simplicity, we believe that SC-BNs could be monolithically integrated using various photonic technologies. For instance, Fig. 2 shows a possible implementation of an 8×8 SC-BN (Omega network) with electro-absorption (EA) gates that we are currently investigating.

In what follows, we assume that the two switches run through a periodic sequence of *N* permutations.

For simplicity, let P^m define the permutation that connects input *i* to output $(i+m)$ mod *N* at the time slots $m+kN$ $(0 \le m,i \le N-1)$. In other words, at time slot *t*, input port *i* is connected to output port *j* where $j=(i+t)$ mod *N*. During a cycle of *N* time slots, each input is connected once to each output. Also, an input is connected to an output every *N* time slots.

2- Single-stage optical buffer

The emulation of a VOQ by an OB is made possible because of the deterministic nature of the TDM switch. It is indeed

possible to forecast at which time slots an input is connected to an output: a packet *p(t,d)* arriving after the load-balancing to port *r* during the time slot *t* will be able to reach its destination *d* at the time slots $\Delta t+kN$ where $\Delta t=(d-r-t)$ mod *N* and *k* is an integer. As Δt and *N* are fixed, only *k* has to be determined by the scheduling algorithm.

In the particular case of the LB-OPS, a VOQ with FIFO queues of length *b* can be emulated by a single-stage OB that is able to delay a packet from *0* to *bN-1* time slots (see Fig. 3a). Clearly, contention can occur at the exit of the OB only for packets with the same destination: two packets $p(t_1,d)$ and $p(t_2,d)$ can collide if $t_1+\Delta t_1+k_1N = t_2+\Delta t_2+k_2N$. To prevent this, the following scheduling is used. To each destination is associated a counter *dest(d)*. It is incremented by *1* each time a packet destined for *d* arrives. Also, at time slots *t'+1,* where *t'* mod *N* = *(d-r)* mod *N*, *dest(d)* is automatically decremented by *1* because a packet destined for *d* has left the OB in the previous time slot.

A packet *p(t,d)* is thus delayed $dest(d).N+\Delta t$ time slots in the $dest(d).N+\Delta t$ fiber delay line (FDL).

For each OB, *N* counters are needed, one for each output.

In practice, a single-stage OB appears to be impractical due to the large number of FDLs/switches required.

2- Double-stage optical buffer

The double-stage OB shown in Fig. 3b is composed of a cycle optical buffer (C-OB), which can delay a packet by multiple cycle periods, and a slot-adjustment optical buffer (SA-OB), which can delay a packet from *0* to *N*-1 time slots. A packet *p(t,d)* requiring a total delay of $\Delta t+kN$ time slots is therefore delayed *kN* time slots in the C-OB, then Δt time slots in the SA-OB.

An additional source of contention can happen in this two-stage configuration at the exit of the C-OB for packets with arrival time intervals proportional to *N*: two packets $p(t_1,d_1)$ and $p(t_2,d_2)$ can collide if $t_1+k_1N = t_2+k_2N$. To solve this issue, another set of *N* counters, *permut(m),* are introduced. A counter *permut(m)* is incremented by *1* if a packet arrives when permutation P^m is set *(t=m+kN)*. Also, at time slots *t'+1,* where *t'* mod *N* = *m*, *permut(m)* is automatically decremented by *1*. Moreover, to insure that no collision occurs at the end of the SA-OB, which is determined by the counter *dest(d)*, both counters should adjusted according to the following rule:

permut(m)= dest(d) := max (permut(m),dest(d)).

This scheduling is not optimal in terms of packet loss probability and average delay but its computational complexity remains *O(1)*.

When a packet *p(t,d)* arrives at the entrance of the OB, it is first delayed by $dest(d).N$ time slots in the *dest(d)* FDL of the C-OB. At the time slot $t+dest(d).N$, it will exit it and will be next delayed Δt time slots in the Δt FDL of the SA-OB. At the time slot $t+\Delta t+dest(d).N$, it will pass through the TDM switch and exit at the output port *d*.

In practice, the OBs can be fabricated from an assembly of one-dimensional gate arrays, optical fibers and passive couplers or wavelength-sensitive devices [2].

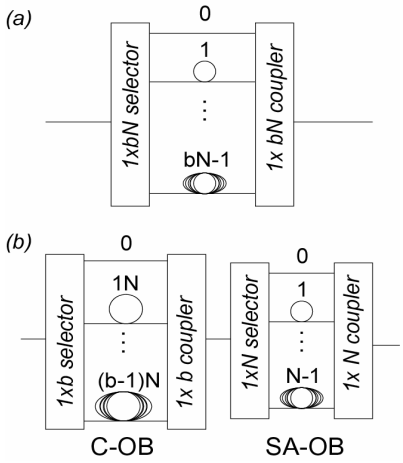

Fig. 3. (a) Single-stage and (b) double-stage optical buffer

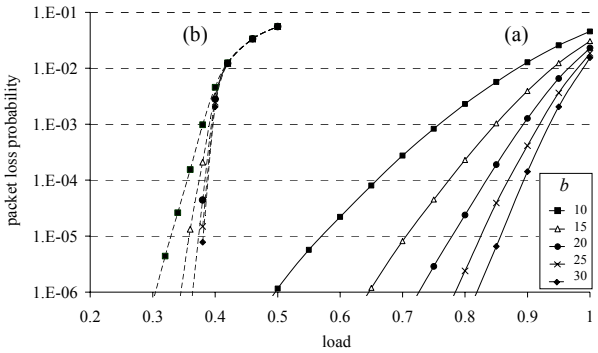

Fig. 4. Packet loss probability for (a) a one-stage OB and (b) a double stage optical buffer.

IV- Simulation

As the function of the load-balancing switch is to uniformly distribute the input traffic, the packet-arrival process at the entrance of the OBs is assumed to be an i.i.d. Bernoulli one. The probability that a packet comes during a time slot is constant and called load. Simulations were carried out for an optical buffer of equivalent FIFO queue length *b* for *N=16*. About 10^8 packets were generated per load (see Fig. 4).

For a buffer size *b =20,* a packet loss probability of 10^{-6} can be obtained for a load less than 0.72 in the case of a single-stage OB and 0.38 in the case of a double-stage one. Also, it can be seen that the double-stage OB saturates above a certain load (*0.42* for *N=16*) whatever the buffer size is.

The performances of the LB-OPS with double-stage OBs can however be improved by adding an extra-balancing switch, what will be discussed at the conference.

V- Conclusion

The implementation of the LB-OPS using stage-controlled Banyan networks and optical buffers illustrates its potential in terms of simplicity and integration ability. In addition, it requires neither a central scheduler nor complex computation.

References

1. I. Keslassy et al., *ACM SIGCOMM* (2003), 189.
2. D.K. Hunter et al., J. Lightwave Technol., 26(1998), 2081.
3. C.S. Chang et al., Comp. Commun., 25 (2002), 623.
4. A. Massini, Discrete Applied Math., 128 (2003), 43.